Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon

Bert Réveil[1], Jean-Pierre Martens[1], Henk van den Heuvel[2]

[1]DSSP group, ELIS, UGent, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
[2]CLST, Fac. of Arts, Radboud Universiteit Nijmegen, The Netherlands
breveil@elis.ugent.be, martens@elis.ugent.be, h.vandenheuvel@let.ru.nl

Abstract

This paper deals with the task of large vocabulary proper name recognition. In order to accomodate a wide diversity of possible name pronunciations (due to non-native name origins or speaker tongues) a multilingual acoustic model is combined with a lexicon comprising 3 grapheme-to-phoneme (G2P) transcriptions (from G2P transcribers for 3 different languages) and up to 4 so-called phoneme-to-phoneme (P2P) transcriptions. The latter are generated with (speaker tongue, name source) specific P2P converters that try to transform a set of baseline name transcriptions into a pool of transcription variants that lie closer to the 'true' name pronunciations. The experimental results show that the generated P2P variants can be employed to improve name recognition, and that the obtained accuracy is comparable to what is achieved with typical (TY) transcriptions (made by a human expert). Furthermore, it is demonstrated that the P2P conversion can best be instantiated from a baseline transcription in the name source language, and that knowledge of the speaker tongue is an important input as well for the P2P transcription process.

1. Introduction

Important applications of automatic speech recognition (ASR), such as voice-driven navigation (GPS) systems and automated call routing, require the recognition of proper names from a large set. The latter still remains a challenging task because of the mismatch that often exists between the way names are treated in the recognition system (by phonemic transcriptions and acoustic models) and the way they are actually pronounced by the user of the system.

Since for the envisaged applications manual transcriptions are too costly to collect (Béchet et al., 2002; Li et al., 2007), the phonemic transcriptions are normally generated by a grapheme-to-phoneme (G2P) converter. However, this converter is usually trained on common text material and therefore not well prepared to deal with archaic name spellings and name parts originating from a foreign language. Furthermore, there is often no standard pronunciation of a particular name (Gao et al., 2001; Li et al., 2007), and if foreign speakers are to be accommodated, the pronunciations can be very accented (Compernolle, 2001; Goronzy et al., 2004).

There are two well-known strategies for tackling the above problems, namely lexical modeling and acoustic modeling/adaptation. *Lexical modeling* aims at employing, for each name, multiple phonemic transcriptions representing expectable pronunciations. These transcriptions can for instance be constructed by taking acoustic evidence into account (Ramabhadran et al., 1998). But they can also originate from extra (e.g. foreign language) G2P converters (Cremelie and ten Bosch, 2001; Maison et al., 2003), or from phonological rules which are applied to an initially available transcription. These rules can in their turn be obtained by means of a knowledge-based (Bonaventura et al., 1998; Bartkova and Jouvet, 2007) or a data-driven approach (Humphries et al., 1996; Amdal et al., 2000; Goronzy et al., 2004). *Acoustic modeling* aims at optimizing the acoustic model parameters through speaker (cluster) adaptation (Mayfield-Tomokiyo, 2000; Raux, 2004) or through model (re)training with accented speech (van Leeuwen and Orr, 1999; Stemmer et al., 2001; Bartkova and Jouvet, 2007).

In this work, a lexical modeling approach is investigated. It relies on so-called phoneme-to-phoneme (P2P) converters (Yang et al., 2006) to turn one initially available name transcription into a set of alternatives. The P2P converters are trained automatically from pairs of initial and auditorily verified (AV) transcriptions for recorded name utterances. Afterwards, they are used to produce alternative transcriptions for names that did not occur in the training phase.

The work presented in this paper builds on recent work (van den Heuvel et al., 2009; Réveil et al., 2009) in which the P2P approach was tested in combination with a state-of-the-art recognizer comprising an acoustic model trained on Dutch native speech. In the present study, we first of all recall the main experimental results and conclusions of our previous work (Sections 2 to 4). Then our approach is tested in combination with a multilingual acoustic model, and an alternative method for constructing the P2P converters (Section 5) is proposed. Furthermore, it is investigated whether the P2P converters are able to capture the type of knowledge that is commonly used by phoneticians to define the typical (TY) transcription of a name.

2. Experimental set-up

All experiments are conducted on the Autonomata Spoken Name Corpus (ASNC) (van den Heuvel et al., 2008), a corpus of spoken name utterances enriched with several broad phonemic transcriptions per name. The recognition is performed by the state-of-the-art commercially available Vocon3200 recognition engine of Nuance (www.nuance.com). The commercially available G2P converters for Dutch, English and French embedded in the Nuance RealSpeak TTS system were used to construct reference lexicons.

2.1. Spoken name utterances

The ASNC contains name utterances of 120 Dutch, 40 English, 20 French, 40 Moroccan and 20 Turkish native speakers. Recordings were made in two regions: Flanders and the Netherlands. Each speaker read 181 names, 69 person names (first name + family name) and 112 geographical names (street names and city names): (1) 120 Dutch names (40 person names and 80 geographical names), (2) 23 English names (7 person names and 16 geographical names), (3) 15 Moroccan person names and (4) either 23 French names (in Flanders) or 23 Turkish names (in the Netherlands) (7 person names and 16 geographical names). There were 10 mutually exclusive name lists per region, 6 of which were read by 16 speakers, the other 4 by only 6. Because of a few overlaps between the Dutch and the Flemish name lists the ASNC contains only 3540 (instead of 20 x 181 = 3620) different names. For all experiments, the corpus was divided in a train set and a test set, and there was *no overlap in speakers nor in name lists* between the two parts.

Although other combinations occur, our former and current work focuses on Dutch natives uttering Dutch and non-native names, and non-native Dutch speakers uttering Dutch names. Table 1 shows the number of utterances in the train and test set for each (speaker tongue, name source) pair. In the following the exemplary combination (DU,FR)

Table 1: Number of available utterances in the ASNC train/test set per (speaker tongue,name source) combination.

	Set	DU	EN	FR	MO	TU
(DU,*)	train	9960	1909	966	1245	943
	test	4440	851	414	555	437
(*,DU)	train	9960	3000	1680	3360	1560
	test	4440	1800	720	1440	840

refers to Dutch speakers reading French names.

Each name in the corpus comes with a typical transcription (TY). It represents, according to a human expert, a valid and likely pronunciation of that name by a native Dutch speaker. Each name token (utterance) comes with an auditorily verified (AV) transcription which is the transliteration of that utterance as it was made by a human expert who could listen as many times as needed to the utterance: see also (van den Heuvel et al., 2008).

2.2. Recognition engine

The Vocon3200 speech recognition engine was delivered with two acoustic models. The first one is a monolingual Dutch acoustic model (AC-MONO), trained on speech of native Dutch speakers from the Netherlands and Flanders. Its underlying phoneme set consists of 45 phonemes. The second model is a multilingual model (AC-MULTI), trained on the same data as AC-MONO, but supplemented with UK English, French and German speech. The Dutch portion only constitutes 20% of the total training data. The underlying phoneme set consists of 80 phonemes, and the model contains roughly 70% more parameters than AC-MONO. Models for phonemes appearing in multiple languages have seen data from all the languages in which they appear.

2.3. Recognition task

The recognizer was operated with a grammar that is just a loop of the 3540 different names appearing in the ASNC. As a performance measure the Name Error Rate (NER) was adopted, defined as the percentage of name utterances that was not, as a whole, recognized correctly.

3. Building P2P converters

In order to learn a P2P converter one considers the orthographic transcription, an *initial* G2P transcription and a *target* phonemic transcription (e.g. the TY or the AV transcription) of a sufficiently large collection of name utterances. The constructed 3-tuples are supplied to an automatic learning process that is visualized in Figure 1.

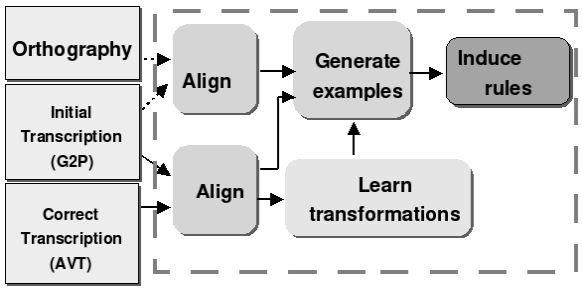

Figure 1: Process for automatically learning of a P2P converter.

The process first retrieves candidate input/output pattern transformations accounting for systematic discrepancies between the aligned initial and target transcriptions. Given these transformations, training examples are constructed at every location where an input pattern of the candidate transformation list is encountered in the processed name. Each example represents (1) the input pattern, also called the *focus*, (2) the linguistic context in which it occurs, and (3) the target output pattern to select: either the input pattern itself (if no difference occurs between initial and target transcription) or an output pattern derived from the candidate transformation list. In order to describe the linguistic context of the focus the following 22 features are being employed: (1) the two phonemes to the left and to the right of the focus, (2) the vowels of the focus syllable, the previous syllable and the next syllable, (3) the stress levels of these three syllables, (4) the initial two characters of the graphemic pattern corresponding to the focus, (5) the graphemes left and right of the former graphemic pattern, (6) a flag which is true if the graphemic pattern corresponding to the focus ends on a dot, (7) the orthographic pattern corresponding to the focus syllable, the previous and the next syllable if they belong to a predefined list of patterns that may induce an error, (8) the word prefix and suffix (if they belong to predefined lists of prefixes and suffixes), and (9) the positions (in syllables) of the focus start/end w.r.t. the word prefix/suffix. To obtain the syllable, prefix and suffix lists needed for the context description, the example generator has to be run twice: in the first run it uses a dummy list and it automatically retrieves

a meaningful list that can then be utilized in the second run. From the training examples, the system finally learns for each candidate transformation a decision tree with a set of stochastic correction rules attached to each of its leaf nodes. In generation mode, the P2P converter parses the aligned initial phonemic transcription from left to right, and by applying rules at different places in the input transcription, it generates multiple pronunciations with attached probabilities. The P2P rules are expected to capture generic knowledge that is applicable to unseen names.

4. Former experimental results

Our former and present experiments focus on native Dutch speakers reading Dutch and foreign names, and on non-native Dutch speakers reading Dutch names. The foreign name source (= language of origin of the name) as well as the foreign speaker tongue (= mother tongue of the non-native speaker) was either English (EN), French (FR), Moroccan (MO) or Turkish (TU). The main conclusions that could be drawn from our former work (with AC-MONO as acoustic model) are:

1. Supplementing the baseline native Dutch G2P transcription of a name with *nativized*[1] transcriptions emerging from two foreign G2Ps (English and French) substantially[2] improves the recognition of foreign names with a name source that is covered by one of these foreign G2Ps. These results clearly support the hypothesis that native speakers use their English and French language knowledge when uttering English and French names. They also comply with the findings of (Cremelie and ten Bosch, 2001; Maison et al., 2003).

2. If in the above lexicon, each name also gets all the AV transcriptions appearing in the ASNC (including those for the name tokens of the test set), the recognition becomes very accurate, in spite of the large number of transcriptions per name.

3. If the three nativized G2P transcriptions of a name are supplemented with transcriptions generated by a (speaker tongue, name source) specific P2P converter that was trained on discrepancies between Dutch G2P transcriptions and the AV transcriptions, there is a substantial gain in the recognition of foreign names spoken by Dutch natives.

4. Combining a multilingual acoustic model with a multilingual lexicon comprising plain (non-nativized) transcriptions of a Dutch, English and French G2P converters - like in (Stemmer et al., 2001; Bartkova and Jouvet, 2007) - substantially improves the recognition of all foreign names uttered by Dutch natives, as well as that of Dutch names uttered by foreign speakers whose mother tongue was present in the acoustic model training data. However, the new multilingual acoustic model also induced a substantial recognition loss for Dutch names read by Dutch natives. These results confirm that non-native speakers use accented sounds which are better modeled by the multilingual acoustic model.

To facilitate the interpretation of the new results presented in the next section, we recall in Table 2 the NERs we obtained with three systems covering different (acoustic model, lexicon) combinations. The figures seem to support the hypothesis that the English and French G2P converters can often produce better transcriptions for Moroccan and Turkish names than the Dutch G2P converter.

Table 2: NER (%) – AC-MONO + Dutch G2P (system 1); AC-MONO + Dutch, nativized English and nativized French G2P (system 2); AC-MULTI + Dutch, English and French G2P (system 3).

	Sys.	**DU**	**EN**	**FR**	**MO**	**TU**
(DU,*)	1	3.5	26.7	13.5	8.3	17.8
	2	3.5	11.4	3.9	7.9	15.8
	3	4.4	7.5	2.7	5.8	12.8
(*,DU)	1		15.8	22.5	17.6	29.6
	2		15.2	20.1	17.4	29.9
	3		10.9	12.9	15.6	30.5

5. Extensions of our former work

In this section we provide a number of new experiments and we assess a new strategy for constructing suitable P2P converters.

5.1. Monolingual versus multilingual transcriptions

By moving from System 2 to System 3 (Table 2) two changes were made, and one may wonder how much of the improvement was due to the change of the acoustic model and how much was due to the change of the phonemic transcriptions. To that end we have also tested AC-MULTI in combination with nativized G2P transcriptions. The figures in Table 3 clearly show that practically all the improvement was due to the change of the acoustic model. This is an important result for us since it implies that the nativized transcriptions provided in the ASNC are appropriate as target transcriptions for P2P learning.

Table 3: NER (%) – AC-MULTI + Dutch, English and French G2P (system 3); AC-MULTI + Dutch, nativized English and nativized French G2P (system 4).

	Sys.	**DU**	**EN**	**FR**	**MO**	**TU**
(DU,*)	3	4.4	7.5	2.7	5.8	12.8
	4	4.3	7.9	2.4	6.7	13.0
(*,DU)	3	4.4	10.9	12.9	15.6	30.5
	4	4.3	10.6	12.8	15.6	29.4

[1] Nativized means that foreign phonemes were mapped to native (Dutch) counterparts using a mapping table that was designed by a human expert.

[2] In this work the term 'substantial' is used whenever a gain or loss of over 20% relative w.r.t. the reference performance is observed.

5.2. Including P2P transcription variants

As in (van den Heuvel et al., 2009), it is assumed that the mother tongue of the speaker and the source language of the name are known, and that one can thus train specific P2P converters for the distinct (DU,*) and (*,DU) combinations. We argue that this is a reasonable claim since the name source language in a car navigation system for instance will greatly depend on the user's location, while the speaker tongue can be deduced from one single control question. Because of this assumption we furthermore decided to evaluate each of the trained P2P converters separately, meaning that the recognition results for System 5 in the 9 cells of Table 4 are obtained with 9 different lexicons, in which only those names for which the P2P converter is intended actually receive additional P2P transcriptions.
All learned P2P converters depart from the Dutch G2P transcription of a name as the initial transcription. In Table 4 we summarize the results obtained by supplementing the lexicon of System 4 (3 nativized G2P transcriptions) with a maximum of four transcriptions generated by the appropriate P2P converter. It is interesting to point out that the probabilities of the P2P variants could not be supplied to the recognition engine, and were therefore not taken into account.

Table 4: NER (%) – AC-MULTI + Dutch, nativized English and nativized French G2P (system 4); AC-MULTI + Dutch, nativized English and nativized French G2P and four P2P transcriptions (system 5).

	Sys.	DU	EN	FR	MO	TU
(DU,*)	4	4.3	7.9	2.4	6.7	13.0
	5	4.2	7.5	2.7	2.7	6.6
(*,DU)	4	4.3	10.6	12.8	15.6	29.4
	5	4.2	9.2	10.0	15.3	26.1

Looking at the (DU,*) combinations, it is clear that the P2P transcriptions substantially improve the recognition of Moroccan and Turkish names read by Dutch speakers. This supports the hypothesis that a number of actual pronunciations of these names can (partly) be explained in terms of systematic modifications of the standard Dutch pronunciations. For the English and French names, Dutch speakers also modify their pronunciation rules, but the knowledge captured by the P2P converter does not outperform the knowledge embedded in the English and French G2P converters. Note that in former experiments with a monolingual acoustic model we did find substantial improvements for English and French names as well. Apparently, by using a multilingual acoustic model that can better handle accented sounds, the P2P variants are not needed anymore to get a sufficient match between the actual utterance and its model (the acoustic model and the pronunciation in the lexicon).
Looking at the (*,DU) combinations, there are modest improvements in the recognition of almost all foreign speaker utterances (except for the Moroccans). This means that these speakers somewhat *Dutchify* their pronunciations of Dutch names (i.e. they take Dutch pronunciation rules into account), and that a P2P converter departing from a Dutch G2P transcription is able to model this to some extent. Here the improvements are larger than the ones we found before in combination with a monolingual acoustic model. We argue that the better transcriptions are now more effective because the acoustic models better match the sounds uttered by the foreign speakers.
Increasing the number of P2P transcriptions per name from four to eight did not further reduce the NER. Apparently, additional transcriptions do not help anymore, but they also don't seem to hurt either.

5.3. P2P variants versus TY transcriptions

An interesting question is whether the P2P converters have captured the kind of knowledge that is commonly used by a phonetician to instantiate a typical Dutch pronunciation of a name. To that end we have conducted an experiment in which the single TY transcription that was supplied with the ASNC was added to the lexicon of System 4. The results obtained with this system are summarized in Table 5.

Table 5: NER (%) – AC-MULTI + Dutch, nativized English and nativized French G2P and four P2P transcriptions (system 5); AC-MULTI + Dutch, nativized English and nativized French G2P and TY transcriptions (system 6).

	Sys.	DU	EN	FR	MO	TU
(DU,*)	5	4.2	7.5	2.7	2.7	6.6
	6	3.4	5.9	2.9	3.6	7.6
(*,DU)	5	4.2	9.2	10.0	15.3	26.1
	6	3.4	9.5	12.2	14.9	27.7

The most important conclusion is that in a cross-lingual setting the automatically generated P2P transcriptions of System 5 generally compete well with the TY transcription which require the intervention of a human expert. For the monolingual (DU,DU) combination however, the TY transcriptions are outstanding. Consequently, it may be a viable approach to investigate whether using the TY transcriptions as targets during P2P development would yield better transcriptions here. However, doing so did not really reduce the NER (4.1% instead of 4.2%).
Since the NERs for systems 6 and 4 are very comparable for the foreign utterances of Dutch names, we can conclude that the TY transcriptions are not very succesful at capturing the pronunciations made by non-native speakers. This is no surprise since a TY transcription is actually intended to represent a typical Dutch pronunciation.

5.4. An alternative P2P training strategy

Since the developed P2P converters failed to substantially improve the recognition of Dutch names spoken by non-native speakers, we tested the hypothesis that this is due to *knowledge transfer*, a concept from the second language learning literature which expresses that foreign speakers use their native language knowledge (You et al., 2005) when reading a non-native name. If this hypothesis holds, it would be a viable option to build for the combinations (EN,DU) and (FR,DU), P2P converters that depart from a nativized English/French G2P transcription respectively.

A second plausible hypothesis is that Dutch speakers use pronunciation rules of the name source language to pronounce foreign names. This is a good reason for taking the foreign G2P transcription as the point of departure for the combinations (DU,EN) and (DU,FR) as well.

The results listed in Table 6 were obtained with P2P converters departing from the nativized foreign G2P transcription in all combinations with EN and FR as foreign languages. The combinations that were affected with respect to System 5 are indicated in italic. The figures do not seem

Table 6: NER (%) – AC-MULTI + Dutch, nativized English and nativized French G2P and four P2P transcriptions (system 5); idem, but the P2P converters departed from foreign initials for (DU,EN), (DU,FR), (EN,DU) and (FR,DU) (system 7)

	Sys.	**DU**	**EN**	**FR**	**MO**	**TU**
(DU,*)	5	4.2	*7.5*	*2.7*	2.7	6.6
	7	4.2	*6.0*	*2.7*	2.7	6.6
(*,DU)	5	4.2	*9.2*	*10.0*	15.3	26.1
	7	4.2	*9.1*	*10.7*	15.3	26.1

to support the hypothesis of knowledge transfer. The hypothesis that Dutch speakers use their name source language knowledge does seem to hold. The fact that it does not for French names can be due to the fact that the NER for (DU,FR) was already quite low, and that the remaining errors correspond to pronunciations which are difficult to model. Another reason might be that the average English knowledge is larger than the average French knowledge of the Flemish speakers.

The above findings are confirmed by a control experiment with System 4. It was investigated (on the training set) which G2P transcription got selected as the top hypothesis whenever the recognition was correct. For the combinations (EN,DU) and (FR,DU) that happened to be the Dutch transcription in 82% and 79% of the cases. For the combinations (DU,EN) and (DU,FR) the nativized English and French G2P transcriptions were chosen in 59% and 79% of the cases respectively.

5.5. Influence of the speaker tongue

Based on the previous section, one may wonder how important it really is to take account of the speaker tongue for the development of the P2P converters. We therefore performed some additional experiments on the (*,DU) combinations, in which they were covered by a single P2P converter. Three such converters were trained: one on all training utterances (System 8), one on only the utterances of the Dutch speakers (System 9), and one on only the utterances of the foreign speakers (System 10). The NERs obtained with the corresponding P2P transcriptions are listed in Table 7. The best results are obtained with System 10, but even for this system the reduction in NER with respect to System 4 for the (foreign speaker,DU) pairs is only about 60% of the reduction that was obtained with System 5. This clearly indicates that knowledge transfer does play a vital role in the pronunciation process after all.

Table 7: NER (%) – AC-MULTI with different lexicons: (1) Dutch + nativized English + nativized French G2P (System 4); (2) Lexicon of system 4 + P2P transcriptions (System 5); (3) Lexicon of system 4 + P2P transcriptions from one P2P trained on (*,DU) (system 8); (4) Lexicon of system 4 + P2P transcriptions from one P2P trained on (DU,DU) (system 9); (5) Lexicon of system 4 + P2P transcriptions from one P2P trained on (foreign,DU) (system 10)

	Sys.	**DU**	**EN**	**FR**	**MO**	**TU**
(*,DU)	4	4.3	10.6	12.8	15.6	29.4
	5	4.2	9.2	10.0	15.3	26.1
	8	4.0	10.1	12.4	15.3	27.7
	9	4.2	10.1	11.9	15.4	28.9
	10	3.9	9.7	11.1	15.3	27.5

A more surprising result is that the P2P converters (partly) trained on foreign speakers outperform the one exclusively trained on Dutch speakers. We argue that the relevant pronunciation variations are in the latter case not picked up by the P2P learning tools, presumably because they occur only occasionally. To verify this hypothesis, we trained a second P2P converter on a subset of the original (DU,DU) training data, namely, the utterances for which recognition with System 4 (3 nativized G2P transcriptions) failed, and an equally large but randomly selected set of utterances for which the recognition was correct. With this P2P converter, we could reduce the NER for (DU,DU) to 3.8%.

6. Conclusions and future work

In this paper it was demonstrated that proper name recognition in a cross-lingual setting (non-native speakers or names) benefits a lot from a multilingual acoustic model and from nativized transcriptions emerging from foreign G2Ps. Furthermore, it was shown that if the mother tongue of the speaker and the source language of the names are known, one can further reduce the error rate by supplementing the lexicon with transcriptions that are generated by a P2P converter that was learned to modify the baseline transcriptions in the direction of the AV transcriptions. The performance gains could be attained on a test set sharing no speakers nor names with the training set.

Unfortunately, the gains are not equally large for all combinations of speaker tongue and name source language. The largest gains are observed for native speakers uttering foreign names emerging from a language that is not covered by the acoustic model training data nor by the available G2P converters. Other significant gains are observed for Dutch names uttered by foreign speakers whose mother tongue is covered by the acoustic model training data.

Since we have not witnessed any sign of increased lexical confusability due to the presence of multiple transcriptions per name in the lexicon, and since this confusability is acknowledged to be important in a common large vocabulary speech recognition system, we plan to conduct a control experiment with a much larger vocabulary size to establish whether this phenomenon is also relevant for the recogni-

tion of relatively long names as appearing in a car navigation or call routing application.

7. Acknowledgements

The presented work was carried out in the context of the Autonomata Too research project, granted under the Dutch-Flemish STEVIN program.

8. References

I. Amdal, F. Korkmazskiy, and A. C. Surendran. 2000. Joint pronunciation modelling of non-native speakers using data-driven methods. *Proceedings ICSLP*, pages 622–625, October.

K. Bartkova and D. Jouvet. 2007. On using units trained on foreign data for improved multiple accent speech recognition. *Speech Communication 49*, pages 836–846.

F. Béchet, R. de Mori, and G. Subsol. 2002. Dynamic generation of proper name pronunciations for directory assistance. *Proceedings ICASSP*, pages 745–748, May.

P. Bonaventura, F. Gallocchio, J. Mari, and G. Micca. 1998. Speech recognition methods for non-native pronunciation variants. *Proceedings ESCA Workshop on Modeling Pronunciation Variation for ASR*, pages 17–22.

D. Van Compernolle. 2001. Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication 35*, pages 71–79, August.

N. Cremelie and L. ten Bosch. 2001. Improving the recognition of foreign names and non-native speech by combining multiple grapheme-to-phoneme converters. *Proceedings ISCA ITRW on Adaptation Methods for Speech Recognition*, pages 151–154.

Y. Gao, B. Ramabhadran, J. Chen, H. Erdoğan, and M. Picheny. 2001. Innovative approaches for large vocabulary name recognition. *Proceedings ICASSP*, pages 53–56, May.

S. Goronzy, S. Rapp, and R. Kompe. 2004. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication 42*, January.

J. J. Humphries, P.C. Woodland, and D. Pearce. 1996. Using accent-specific pronunciation modelling for robust speech recognition. *Proceedings ICSLP*, pages 2324–2327, October.

X. Li, A. Gunawardana, and A. Acero. 2007. Adapting grapheme-to-phoneme conversion for name recognition. *Proceedings ASRU*, December.

B. Maison, S. Chen, and P. Cohen. 2003. Pronunciation modeling for names of foreign origin. *Proceedings ASRU*, pages 429–434.

L. Mayfield-Tomokiyo. 2000. Lexical and acoustic modeling of non-native speech in LVCSR. *Proceedings ICSLP*, pages 346–349, October.

B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan. 1998. Acoustics-only based automatic phonetic baseform generation. *Proceedings ICASSP*, pages 309–312, May.

A. Raux. 2004. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. *Proceedings Interspeech*, pages 613–616, October.

B. Réveil, J.-P. Martens, and B. D'Hoore. 2009. How speaker tongue and name source language affect the automatic recognition of spoken names. *Proceedings Interspeech*, pages 2995–2998, September.

G. Stemmer, E. Nöth, and H. Niemann. 2001. Acoustic modeling of foreign words in a german speech recognition system. *Proceedings Eurospeech*, pages 2745–2748.

H. van den Heuvel, J.-P. Martens, B. D'Hoore, C. D'Haene, and N. Konings. 2008. The Autonomata Spoken Name Corpus. *Proceedings LREC*.

H. van den Heuvel, B. Réveil, and J.-P. Martens. 2009. Pronunciation-based asr for names. *Proceedings Interspeech*, pages 2991–2994, September.

D. van Leeuwen and R. Orr. 1999. Speech recognition of non-native speech using native and non-native acoustic models. *Proceedings Workshop MIST*.

Q. Yang, J.-P. Martens, N. Konings, and H. van den Heuvel. 2006. Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names. *Proceedings LREC*, pages 287–292.

H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan. 2005. Pronunciation variations of spanish-accented english spoken by young children. *Proceedings Interspeech*, pages 749–752, September.